US 20080167874A1

(19) United States

(12) Patent Application Publication

(10) Pub. No.: US 2008/0167874 A1

(43) Pub. Date: Jul. 10, 2008

Eide et al.

(54) METHODS AND APPARATUS FOR MASKING LATENCY IN TEXT-TO-SPEECH SYSTEMS

(76) Inventors: **Ellen Marie Eide**, Tarrytown, NY (US); **Wael Mohamed Hamza**, Yorktown Heights, NY (US)

Correspondence Address:
Robert W. Griffith
RYAN, MASON & LEWIS, LLP
90 Forsest Avenue
Locust Valley, NY 11560

(21) Appl. No.: **11/620,842**

(22) Filed: **Jan. 8, 2007**

Publication Classification

(51) **Int. Cl.**
G10L 15/18 (2006.01)

(52) **U.S. Cl.** **704/257**

(57) **ABSTRACT**

A technique for masking latency in an automatic dialog system is provided. A communication is received from a user at the automatic dialog system. The communication is processed in the automatic dialog system to provide a response. At least one transitional message is provided to the user from the automatic dialog system while processing the communication. A response is provided to the user from the automatic dialog system in accordance with the received communication from the user.

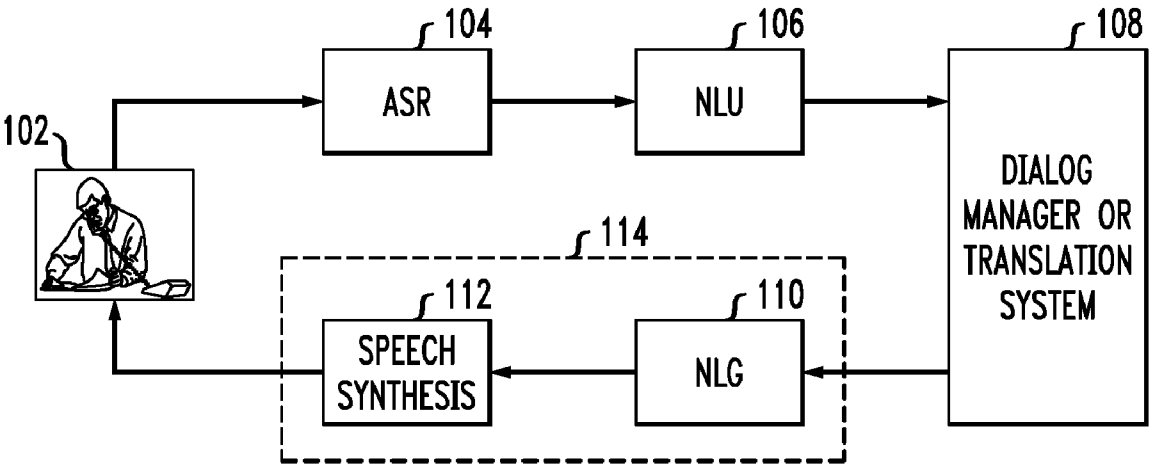


FIG. 1

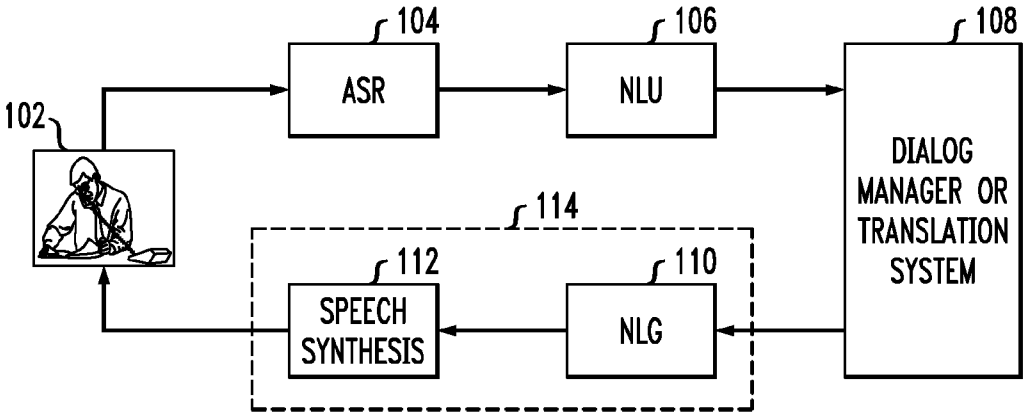

102
104
ASR
106
NLU
108
DIALOG MANAGER OR TRANSLATION SYSTEM
114
112
SPEECH SYNTHESIS
110
NLG

FIG. 2

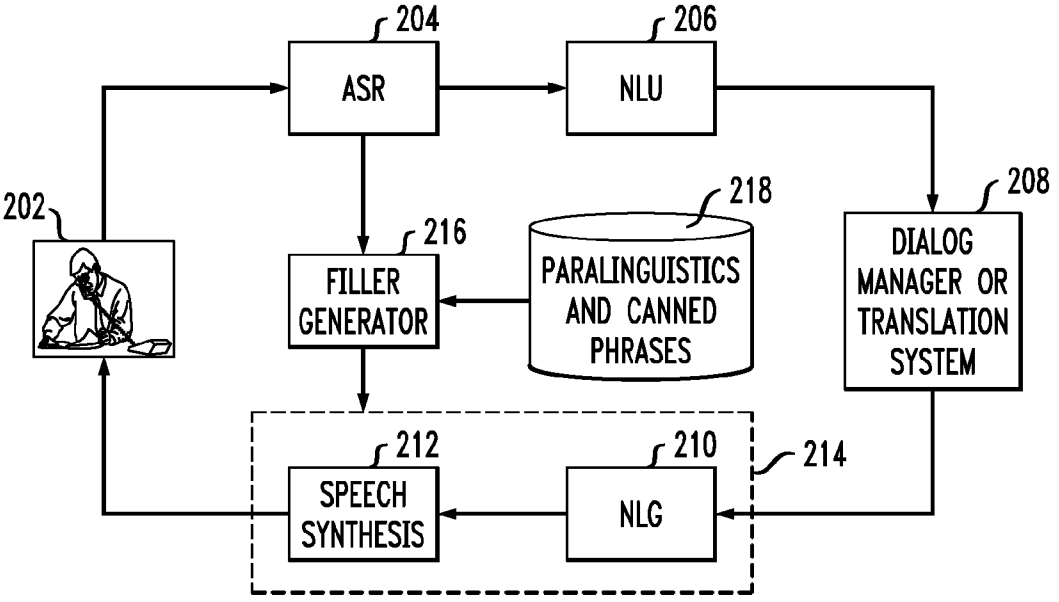

202
204
ASR
206
NLU
208
DIALOG MANAGER OR TRANSLATION SYSTEM
218
PARALINGUISTICS AND CANNED PHRASES
216
FILLER GENERATOR
212
SPEECH SYNTHESIS
210
NLG
214

FIG. 3

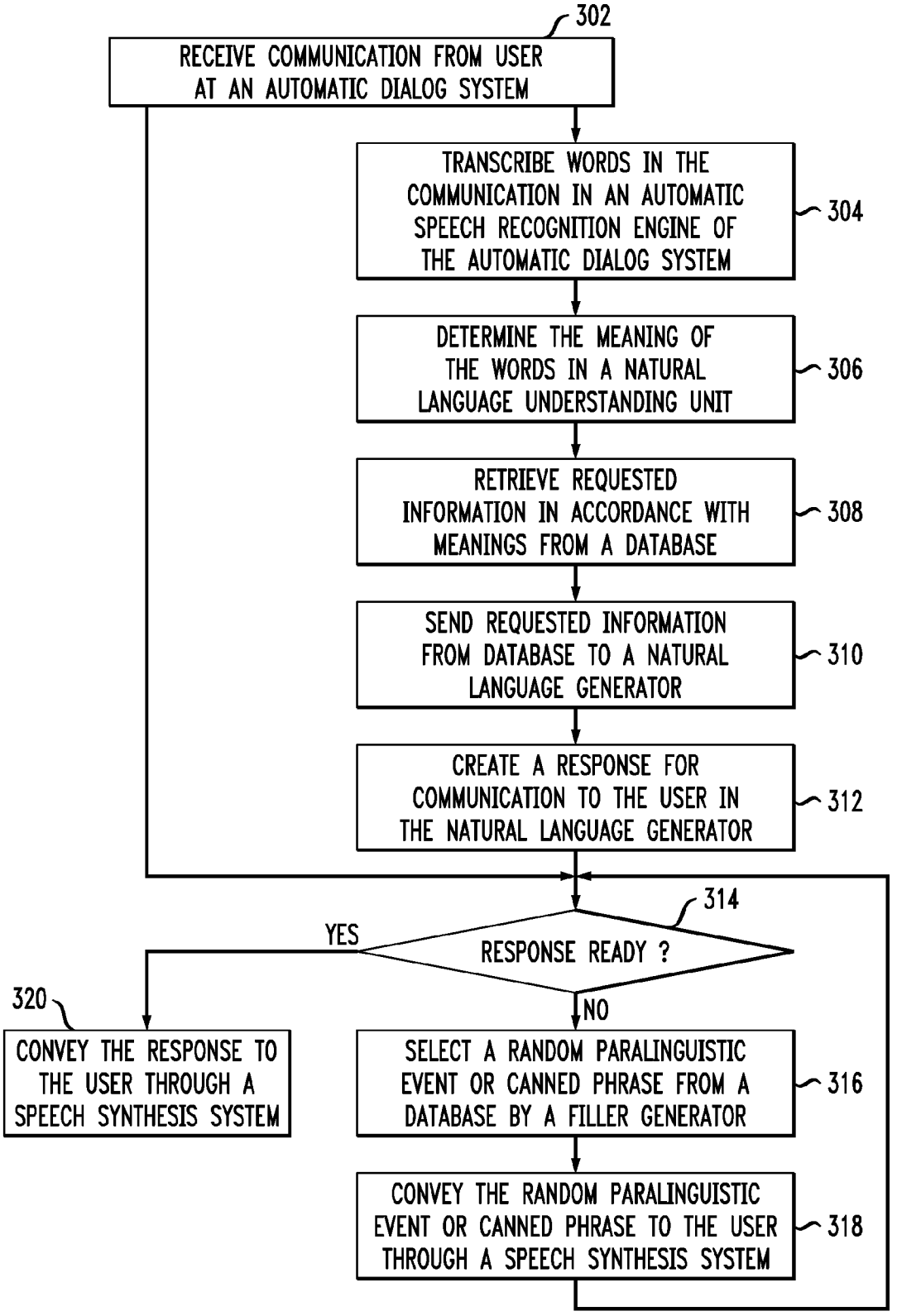
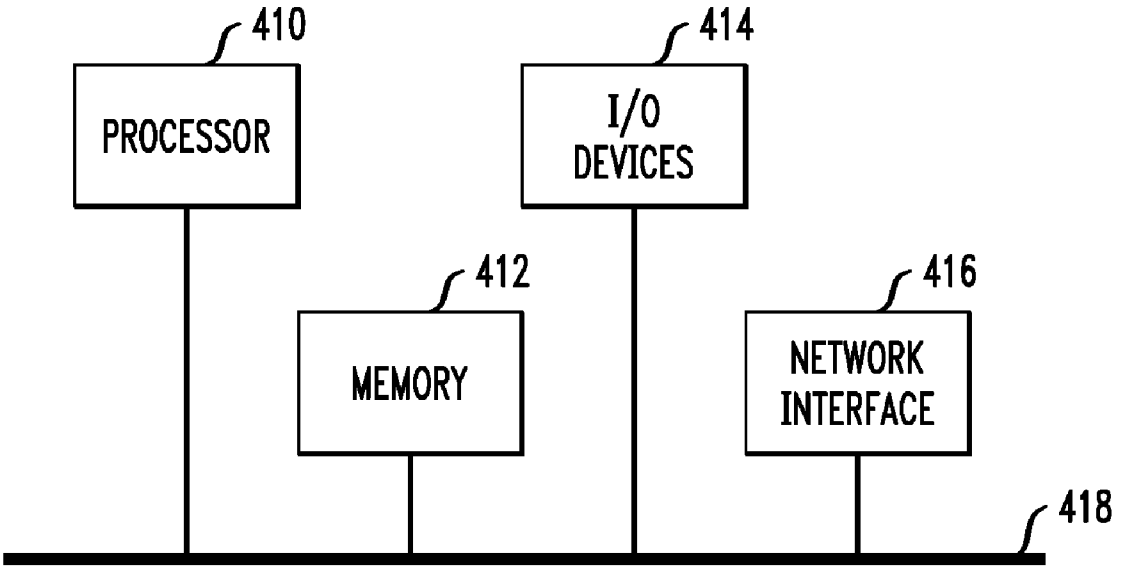

302
RECEIVE COMMUNICATION FROM USER AT AN AUTOMATIC DIALOG SYSTEM
TRANSCRIBE WORDS IN THE COMMUNICATION IN AN AUTOMATIC SPEECH RECOGNITION ENGINE OF THE AUTOMATIC DIALOG SYSTEM
304
DETERMINE THE MEANING OF THE WORDS IN A NATURAL LANGUAGE UNDERSTANDING UNIT
306
RETRIEVE REQUESTED INFORMATION IN ACCORDANCE WITH MEANINGS FROM A DATABASE
308
SEND REQUESTED INFORMATION FROM DATABASE TO A NATURAL LANGUAGE GENERATOR
310
CREATE A RESPONSE FOR COMMUNICATION TO THE USER IN THE NATURAL LANGUAGE GENERATOR
312
314
YES
RESPONSE READY ?
NO
320
CONVEY THE RESPONSE TO THE USER THROUGH A SPEECH SYNTHESIS SYSTEM
SELECT A RANDOM PARALINGUISTIC EVENT OR CANNED PHRASE FROM A DATABASE BY A FILLER GENERATOR
316
CONVEY THE RANDOM PARALINGUISTIC EVENT OR CANNED PHRASE TO THE USER THROUGH A SPEECH SYNTHESIS SYSTEM
318

FIG. 4


410
PROCESSOR
414
I/O
DEVICES
412
MEMORY
416
NETWORK
INTERFACE
418

METHODS AND APPARATUS FOR MASKING LATENCY IN TEXT-TO-SPEECH SYSTEMS

FIELD OF THE INVENTION

[0001] The present invention relates to automatic dialog systems, and more specifically, to methods and apparatus for masking latency in an automatic dialog system.

BACKGROUND OF THE INVENTION

[0002] In telephony applications, text-to-speech (TTS) systems may be utilized in the production of speech output as part of an automatic dialog system. Typically during a call session, automatic dialog systems first transcribe the words communicated by a caller through an automatic speech recognition (ASR) engine. A natural language understanding (NLU) unit in communication with the speech recognition engine is used to uncover the meanings of the caller's words. These meanings may then be interpreted to determine requested information, which may be retrieved from a database by a dialog manager. The retrieved information is passed to a natural language generation (NLG) block, which forms a sentence in response to the caller. The sentence is then output, or spoken, to the caller through a TTS speech synthesis system.

[0003] A TTS system may be utilized in many current real world applications as a part of an automatic dialog system. For example, a caller to an air travel system may communicate with a TTS system to receive air travel information, such as reservations, confirmations, schedules, etc., in the form of TTS generated speech.

[0004] The information passed from the NLG to the TTS speech synthesis system is fed in a time-critical fashion. Unfortunately, the output incurs a compounded latency comprising the processing latencies of the ASR, NLU and NLG processors. Delays between the end of the caller's statement and the output, or spoken reply to the caller, may lead to confusion or frustration on the part of the caller.

[0005] Typically, delays or latencies are masked by playing "earcons", such as, for example, music. Such earcons inform the caller that the system is processing. However, the caller may find the earcons annoying or unnatural.

[0006] Therefore, it is desirable for an automatic dialog system to act similar to a human speaker by masking latency in a more natural manner that does not confuse, frustrate or annoy the caller.

SUMMARY OF THE INVENTION

[0007] The present invention provides techniques for masking latency in an automatic dialog system in a more natural manner by using paralinguistic events or fixed phrases.

[0008] For example, in one aspect of the invention, a technique for masking latency in an automatic dialog system is provided. A communication is received from a user at the automatic dialog system. The communication is processed in the automatic dialog system to provide a response. At least one transitional message is provided to the user from the automatic dialog system while processing the communication. A response is provided to the user from the automatic dialog system in accordance with the received communication from the user.

[0009] In additional embodiments of the present invention, an automatic speech recognition engine of the automatic dialog system may transcribe words in the communication from the user. The meanings of the words may be determined through a natural language understanding unit in communication with the automatic speech recognition engine in the automatic dialog system. Requested information may be retrieved in accordance with the meaning of the words from a database in communication with the natural language understanding unit in the automatic dialog system. The requested information may be sent from the database to the text-to-speech system. The response may be created in a natural language generator of the text-to-speech system. The response may be conveyed to the user through a speech synthesis system of the text-to-speech system, in communication with the natural language generator.

[0010] In a further embodiment of the present invention, in providing the transitional message, a filler generator may select a random message from a database. The random message may be conveyed to the user through a speech synthesis system of the text-to-speech system. Transitional messages may be provided to the user until the response is ready to be provided to the user. In addition, the transitional messages may comprise at least one of a paralinguistic event and a phrase.

[0011] In an additional aspect of the present invention, an automatic dialog system for producing speech output is provided. The automatic dialog system comprises a speech synthesis system that provides at least one transitional message to the user while processing a received communication from the user. The speech synthesis further provides at least one response to the user in accordance with the received communication from the user.

[0012] In an additional embodiment of the present invention, the automatic dialog system may further comprise an automatic speech recognition engine, a natural language understanding unit, a dialog manager, a natural language generator and a filler generator.

[0013] These and other objects, features, and advantages of the present invention will become apparent from the following detailed description of illustrative embodiments thereof, which is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] FIG. **1** is a detailed block diagram illustrating a text-to-speech system utilized in an automatic dialog system;

[0015] FIG. **2** is a detailed block diagram illustrating a text-to-speech system utilized in an automatic dialog system, according to an embodiment of the present invention;

[0016] FIG. **3** is a flow diagram illustrating a latency masking methodology in an automatic dialog system, according to an embodiment of the present invention; and

[0017] FIG. **4** is a block diagram illustrating a hardware implementation of a computing system in accordance with which one or more components/methodologies of the invention may be implemented, according to an embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0018] As will be illustrated in detail below, the present invention introduces techniques for masking latency in an automatic dialog system that is accrued through processing a

response to a user's speech with the use of a filler generator capable of outputting paralinguistic events and fixed phrases.

[0019] Referring initially to FIG. **1**, a detailed block diagram illustrates a TTS system utilized in an automatic dialog system. A caller **102** initiates communication with the automatic dialog system, through a spoken message or request. An ASR engine **104** receives the sounds sent by caller **102** and associates them with words, thereby recognizing the speech of caller **102**. The words are sent from ASR engine **104** to an NLU unit **106**, which determines the meanings behind the words of caller **102**. These meanings are used to determine what information is desired by caller **102**. A dialog manager **108** in communication with NLU unit **106** retrieves the information requested by caller **102** from a database. Dialog manager **108** may also be implemented as a translation system.

[0020] The retrieved information is sent from dialog manager **108** to an NLG block **110**, which forms a message in response to communication from caller **102**, having the requested information. Once the sentence is formed, a speech synthesis system **112**, plays or outputs the sentence to the caller with the requested information. NLG block **110** and speech synthesis system **112** may be considered a TTS system **114** of the automatic dialog system. While generating a response to caller **102**, a latency results that is equal to the sum of the processing latencies of ASR engine **104**, NLU unit **106** and NLG block **110**.

[0021] Referring now to FIG. **2**, a detailed block diagram illustrates a TTS system utilized in an automatic dialog system, according to an embodiment of the present invention. A caller **202** initiates communication with the automatic dialog system, through a spoken message or request. An ASR engine **204** receives the sounds sent by caller **202** and associates them with words, thereby recognizing the speech of caller **202**. The words are sent from ASR engine **204** to an NLU unit **206**, which determines the meanings behind the words of caller **202**. These meanings are used to determine what information is desired by caller **202**. A dialog manager **208** in communication with NLU unit **206** retrieves the information requested by caller **202** from a database. Dialog manager **208** may also be implemented as a translation system. The retrieved information is sent from dialog manager **208** to an NLG block **210**, which forms a message in response to communication from caller **202**, having the requested information.

[0022] As described above, as ASR engine **204**, NLU unit **206** and NLG block **110** are each processing, a latency results that is equal to the sum of the processing latencies of ASR engine **204**, NLU unit **206** and NLG block **210**. To mask the resulting latency, ASR engine **204** first signals a filler generator **216** when caller **202** has finished speaking. Filler generator **216** selects a paralinguistic event or canned/fixed phrase from database **218**. A speech synthesis system **212** of a TTS system **214** may immediately output or play the paralinguistic event or canned phrase from database **218**, or filler generator **216** may delay the output by a few milliseconds before sending the paralinguistic event or canned phrase to speech synthesis system **212**. Filler generator **216** may repeat selecting additional paralinguistic events or canned phrases from database **218** to be output by speech synthesis system **212** until NLG block **210** completes the formation of a response. Once NLG block **210** completes the formation of a response to caller **202**, filler generator **216** stops selecting paralinguistic events and canned phrases to be output, and speech synthesis system **212** plays or outputs the response formed by NLG block **210** to caller **202**.

[0023] The paralinguistic events or canned phrases may be prerecorded into database **218**. The paralinguistic events may be selected randomly and may consist of coughs, breaths, and filled pauses such as, "uh," "um," and "hmmm." Similarly, fixed phrases such as "well . . . " or "let's see . . . " may also be prerecorded into database **200**.

[0024] Referring now to FIG. **3**, a flow diagram illustrates a latency masking methodology in an automatic dialog system, according to an embodiment of the present invention. The methodology begins in block **302**, where an incoming communication is received from a user at an automatic dialog system. Typically a user of an automatic dialog system is a caller attempting to obtain specific information. In block **304**, words in the communication from the user to the automatic dialog system are transcribed in an ASR engine of the automatic dialog system. In block **306**, the meanings of these words are determined through an NLU unit in communication with the ASR engine in the automatic dialog system. In block **308**, information is retrieved from a database in accordance with the meanings of the words. The information is typically that which is sought by the user or caller from the automatic dialog system. The dialog manager of the database is in communication with the NLU unit in the automatic dialog system. In block **310**, the requested information is sent from the database to an NLG. In block **312**, a response containing the requested information is created in the NLG for communication to the caller.

[0025] As the ASR engine, NLU unit, and NLG are processing, a latency results that is equal to a sum of the processing latencies of the ASR engine, NLU unit and NLG. In block **314**, latency is determined by testing whether a response is ready after receiving a communication from a user in block **302**. If a response is not ready, a filler generator selects a paralinguistic event or canned phrase from a database in block **316**. In block **318**, the random paralinguistic event or fixed phrase is conveyed to the user through a speech synthesis system. The methodology then returns to block **314** to determine whether the natural language generator has created the response. If it is determined in block **314** that the response from block **312** is ready, the response is conveyed to the user through the speech synthesis system in communication with the NLG, in block **320**, terminating the methodology.

[0026] While the example has illustrated a telephone-based system, the invention is easily applied in other scenarios such as kiosks and Internet-based applications. Additional embodiments of the present invention may include different automatic dialog system and TTS system components and configurations. The invention may be implemented in any system in which it is desirable to adapt output speech in accordance with the context of the communication.

[0027] Referring now to FIG. **4**, a block diagram illustrates an illustrative hardware implementation of a computing system in accordance with which one or more components/methodologies of the invention (e.g., components/methodologies described in the context of FIGS. **1-3**) may be implemented, according to an embodiment of the present invention. For instance, such a computing system in FIG. **4** may implement the automatic dialog system and the executing program of FIGS. **1-3**.

[0028] As shown, the computer system may be implemented in accordance with a processor **410**, a memory **412**,

I/O devices **414**, and a network interface **416**, coupled via a computer bus **41 8** or alternate connection arrangement.

[0029] It is to be appreciated that the term "processor" as used herein is intended to include any processing device, such as, for example, one that includes a CPU (central processing unit) and/or other processing circuitry. It is also to be understood that the term "processor" may refer to more than one processing device and that various elements associated with a processing device may be shared by other processing devices.

[0030] The term "memory" as used herein is intended to include memory associated with a processor or CPU, such as, for example, RAM, ROM, a fixed memory device (e.g., hard drive), a removable memory device (e.g., diskette), flash memory, etc.

[0031] In addition, the phrase "input/output devices" or "I/O devices" as used herein is intended to include, for example, one or more input devices for entering speech or text into the processing unit, and/or one or more output devices for outputting speech associated with the processing unit. The user input speech and the TTS system annotated output speech may be provided in accordance with one or more of the I/O devices.

[0032] Still further, the phrase "network interface" as used herein is intended to include, for example, one or more transceivers to permit the computer system to communicate with another computer system via an appropriate communications protocol.

[0033] Software components including instructions or code for performing the methodologies described herein may be stored in one or more of the associated memory devices (e.g., ROM, fixed or removable memory) and, when ready to be utilized, loaded in part or in whole (e.g., into RAM) and executed by a CPU.

[0034] Although illustrative embodiments of the present invention have been described herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to those precise embodiments, and that various other changes and modifications may be made by one skilled in the art without departing from the scope or spirit of the invention.

What is claimed is:

1. A method of masking latency in an automatic dialog system, comprising the steps of:

receiving a communication from a user at the automatic dialog system;

processing the communication in the automatic dialog system to provide a response;

providing at least one transitional message to the user from the automatic dialog system while processing the communication; and

providing a response to the user from the automatic dialog system in accordance with the received communication from the user.

2. The method of claim **1**, wherein the step of processing the communication comprises the steps of:

transcribing one or more words in the communication from the user in an automatic speech recognition engine of the automatic dialog system;

determining at least one meaning of the one or more words through a natural language understanding unit in communication with the automatic speech recognition engine in the automatic dialog system;

retrieving requested information in accordance with the at least one meaning of the one or more words from a database in communication with the natural language understanding unit in the automatic dialog system; and

sending the requested information from the database to a text-to-speech system of the automatic dialog system.

3. The method of claim **1**, wherein the step of providing at least one transitional message comprises the steps of:

selecting at least one transitional message from a database by a filler generator of the automatic dialog system; and

conveying the at least one random transitional message to the user through a text-to-speech system of the automatic dialog system.

4. The method of claim **3**, wherein, in the step of selecting at least one transitional message, the at least one transitional message is selected randomly.

5. The method of claim **1**, wherein, in the step of providing at least one transitional message, the at least one transitional message is provided after a specified delay after receiving the communication.

6. The method of claim **1**, wherein, in the step of providing at least one transitional message to the user, the at least one transitional message comprises a plurality of messages.

7. The method of claim **1**, wherein the at least one transitional message comprises at least one of a paralinguistic event and a phrase.

8. The method of claim **1**, wherein the step of providing a response to the user comprises the steps of:

creating the response in a natural language generator of a text-to-speech system of the automatic dialog system; and

conveying the response to the user through a speech synthesis system of the text-to-speech system, in communication with the natural language generator.

9. An automatic dialog system for producing speech output, comprising a speech synthesis system that provides at least one transitional message to the user while processing a received communication from the user and that provides at least one response to the user in accordance with the received communication from the user.

10. The automatic dialog system of claim **9**, wherein the automatic dialog system further comprises:

an automatic speech recognition engine that transcribes one or more words in the received communication;

a natural language understanding unit in communication with the automatic speech recognition engine for determining at least one meaning of one or more words;

a dialog manager in communication with the natural language understanding unit that retrieves requested information from a database in accordance with the at least one meaning of the one or more words;

a natural language generator in communication with the dialog manager for creating the at least one response for the speech synthesis system.

11. The automatic dialog system of claim **9**, further comprising a filler generator that selects the at least one transitional message from a database and conveys the at least one message to the speech synthesis system.

12. The automatic dialog system of claim **11**, wherein the filler generator provides a plurality of transitional messages to the speech synthesis system for the user until the speech synthesis system provides the at least one response to the user.

13. The automatic dialog system of claim **9**, wherein the at least one transitional message comprises at least one of a paralinguistic event and a phrase.

14. Apparatus for producing speech output in an automatic dialog system, comprising:

a memory; and

at least one processor coupled to the memory and operative to: (i) receive a communication from a user at the automatic dialog system; (ii) process the communication in the automatic dialog system to provide a response; (iii) provide at least one transitional message to the user from the automatic dialog system while processing the communication; and (iv) provide a response to the user from the automatic dialog system in accordance with the received communication from the user.

15. The apparatus of claim **14**, wherein the operation of processing the communication comprises the steps of;

transcribing one or more words in the communication from the user in an automatic speech recognition engine of the automatic dialog system;

determining at least one meaning of the one or more words through a natural language understanding unit in communication with the automatic speech recognition engine in the automatic dialog system;

retrieving requested information in accordance with the at least one meaning of the one or more words from a database in communication with the natural language understanding unit in the automatic dialog system; and

sending the requested information from the database to a text-to-speech system of the automatic dialog system.

16. The apparatus of claim **14**, wherein the operation of providing at least one transitional message comprises the steps of:

selecting at least one transitional message from a database by a filler generator of the automatic dialog system; and

conveying the at least one random transitional message to the user through a text-to-speech system of the automatic dialog system.

17. The apparatus of claim **14**, wherein, in the operation of providing at least one transitional message to the user, the at least one transitional message comprises a plurality of messages.

18. The apparatus of claim **14**, wherein the at least one transitional message comprises at least one of a paralinguistic event and a phrase.

19. The apparatus of claim **14**, wherein the operation of providing a response to the user comprises the steps of:

creating the response in a natural language generator of a text-to-speech system of the automatic dialog system; and

conveying the response to the user through a speech synthesis system of the text-to-speech system, in communication with the natural language generator.

20. An article of manufacture for producing speech output in an automatic dialog system, comprising a machine readable medium containing one or more programs which when executed implement the steps of:

receiving a communication from a user at the automatic dialog system;

processing the communication in the automatic dialog system to provide a response;

providing at least one transitional message to the user from the automatic dialog system while processing the communication; and

providing a response to the user from the automatic dialog system in accordance with the received communication from the user.

* * * * *